The Networking and information Technology Research and Development (NITRD) Program
Faster Administration of Science and Technology Education and Research (FASTER) Community of Practice

Data Sharing and Metadata Curation: Obstacles and Strategies

Transdisciplinary Community Perspectives: Earth Cube

Joel Cutcher-Gershenfeld,
University of Illinois and WayMark Systems

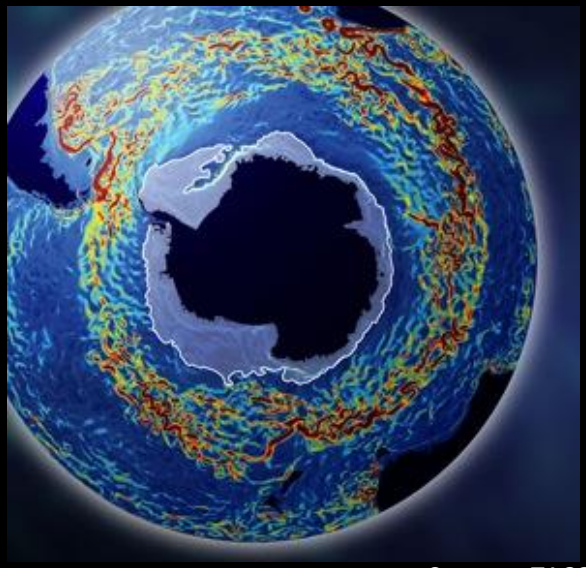
Support from the National Science Foundation is deeply appreciated:
NSF-VOSS EAGER 0956472, "Stakeholder Alignment in Socio-Technical Systems,"
NSF OCI RAPID 1229928, "Stakeholder Alignment for EarthCube,"
NSF SciSPR-STS-OCI-INSPIRE 1249607, "Enabling Transformation in the Social Sciences, Geosciences, and Cyberinfrastructure,"
NSF I-CORPS 1313562 "Stakeholder Alignment for Public-Private Partnerships"

WayMark systems

Three conversations . . .

Source : adapted from 123rf

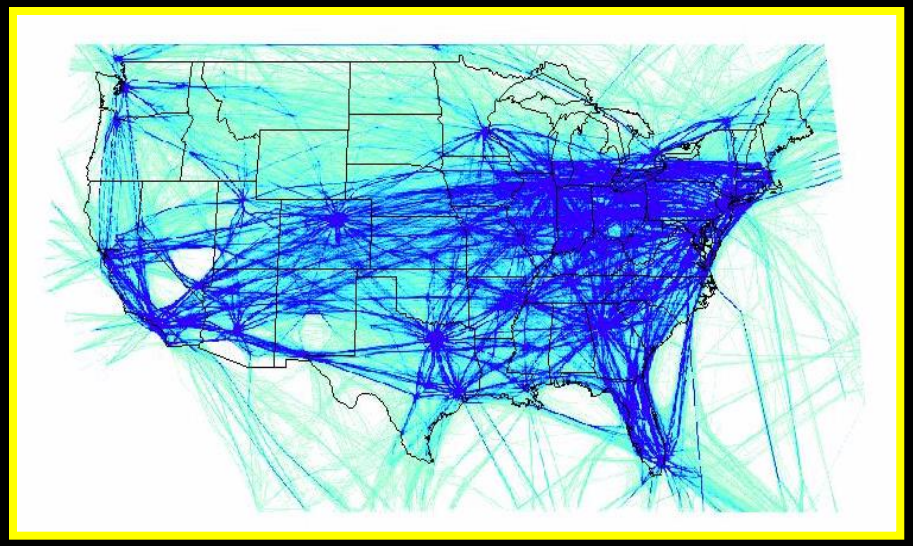
Source : UIUC



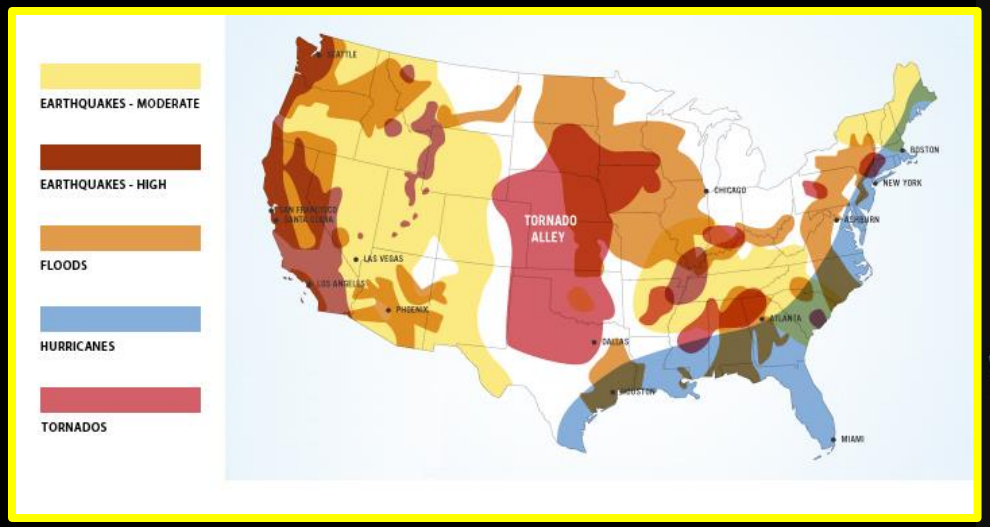
Source : TACC

Institutions ≠ Systems

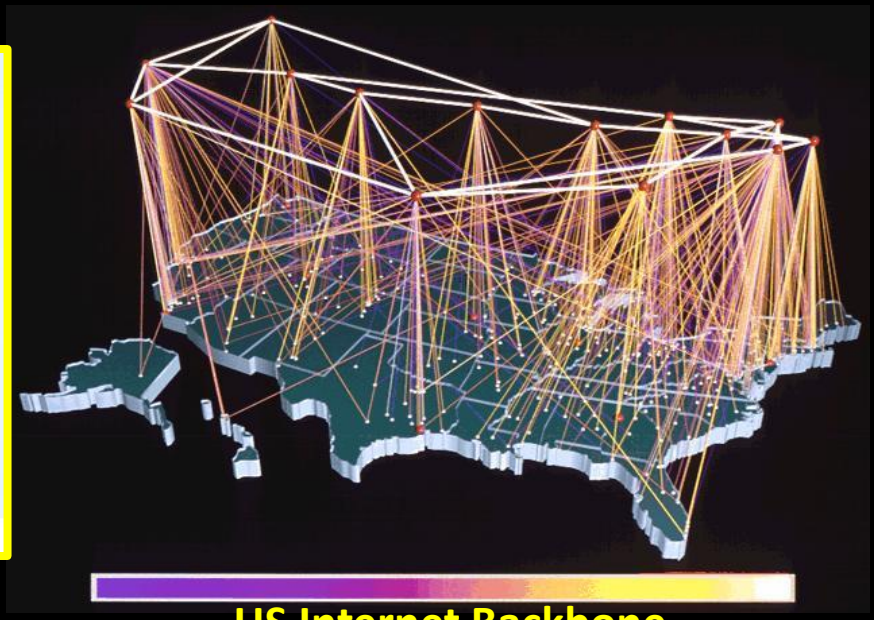
US Power Grid



US Passenger Air Transportation System



Natural Disasters



US Internet Backbone

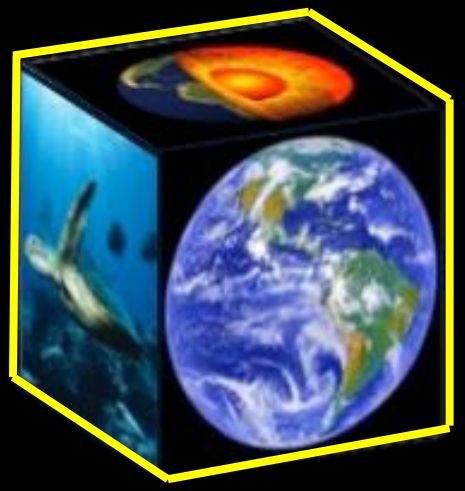
http://www.xprt.net/~rolfsky/internetSite/internet.html

Sources: Carolos A. Osario, ESD Doctoral Seminar, 2004, and Joel Cutcher-Gershenfeld

Plus the challenge of accelerating rates of technological change . . .

The Babysitter of the Future

(courtesy of Steve Diggs, Scripps Institution)



"Over the next decade, the geosciences community commits to developing a framework to understand and predict responses of the Earth as a system—from the space-atmosphere boundary to the core, including the influences of humans and ecosystems."

– *GEO Vision report of NSF Geoscience Directorate Advisory Committee, 2009*

1. What metadata, and what kinds of metadata management, are needed to enable re-use of data, both across domains and across silos within domains?

Specify Stakeholders/Domains and Identify Interests

- Atmospheric or Space Weather scientist
- Oceanographer
- Geologist
- Geophysicist
- Hydrologist
- Critical zone scientist
- Climate scientist
- Biologist or Ecosystems scientist
- Geographers
- Computer or Cyberinfrastructure scientist
- Social scientist (Anthropologist, Economist, Psychologist, Sociologist, etc.)
- Other scientist

- Data manager
- High performance computing expert
- Software engineer
- IT user support personnel
- K-12 educator
- Designer/developer of geoscience instrumentation
- Environmental resource manager (e.g. local, state, or federal)
- Other

50+ interest questions, covering:

- Access and Utilization of Data, Observations, Visualizations, and Models – Current State and Desired State
- Increasing Uniformity and Interoperability through the EarthCube Process
- The Scope of the EarthCube Mission
- Stakeholder Relations and Governance
- Your Potential Engagement with EarthCube

Data collection n=1,211

1.	EarthCube Website (n=127)	Mar.-June, 2012
2.	Data Centers (n=578)	Mar.-June 2012
3.	Early Career (n=37)	Oct. 17-18, 2012
4.	Structure and Tectonics (n=24)	Nov. 19-20, 2012
5.	EarthScope (n=22)	Nov. 29-30, 2012
6.	Experimental Stratigraphy (n=21)	Dec. 11-12, 2012
7.	Atmospheric Modeling (n=29)	Dec. 19, 2012
8.	OGC (n=14)	Jan. 13, 2013
9.	Data Assimilation & Ensemble Prediction	Jan. 18, 2013
10.	Critical Zone (n=39)	Jan. 21-23, 2013
11.	Envisioning a Digital Crust (n=23)	Jan. 29-31, 2013
12.	Paleogeoscience (n=40)	Feb. 3-5, 2013
13.	Education & Workforce Training (n=33)	Mar. 3-5, 2013
14.	Petrology & Geochemistry (n=59)	Mar. 6-7, 2013
15.	Sedimentary Geology (n=50)	Mar. 25-27, 2013
16.	Community Geodynamic Modeling (n=42)	Apr. 22-24, 2013
17.	Integrating Inland Waters, Geochemistry, Biogeochem and Fluvial Sedimentology Communities (n=35)	Apr. 24-26, 2013

Note: Some additional respondents from EC website after June are in overall totals.

Hundreds of specific areas of expertise. . .

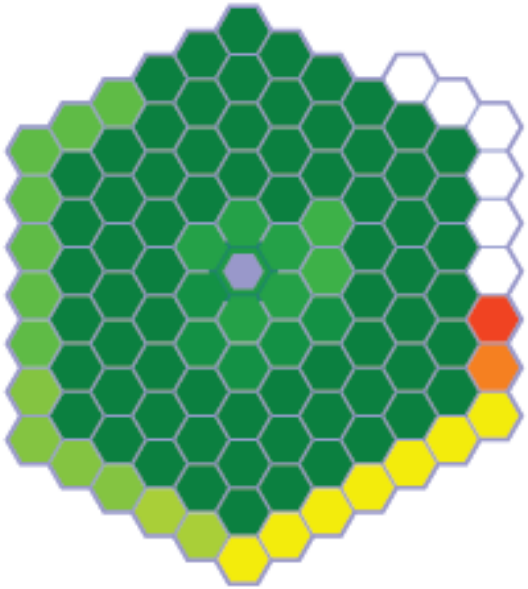
- Air Sea Interaction
- Atmospheric Radiation
- Basalt geochemistry
- Biodiversity Information Networks
- Carbonate Stratigraphy
- Chemical Oceanography
- Coastal Geomorphology
- Computational Geodynamics
- Cryosphere-Climate Interaction
- Disaster Assessment
- Ensemble data assimilation
- Geochronology
- Geoinformatics
- Geomicrobiology
- Glaciology
- Heliophysics
- Isotope Geochemistry
- "It's complicated"
- Magnetospheric Physics
- Mesoscale Meteorology
- Multibeam Bathymetric Data
- Nearshore Coastal Modeling
- Paleoceanography
- Paleomagnetism
- Permafrost Geophysics
- Planetology
- Riverine carbon and nutrient biogeochemistry
- Satellite gravity and altimetry data processing
- Tectonophysics
- Thermospheric Physics
- Watershed Management

IMPORTANCE of integrating multiple datasets, observations, visualization tools, and/or models in your field or discipline

Early Career

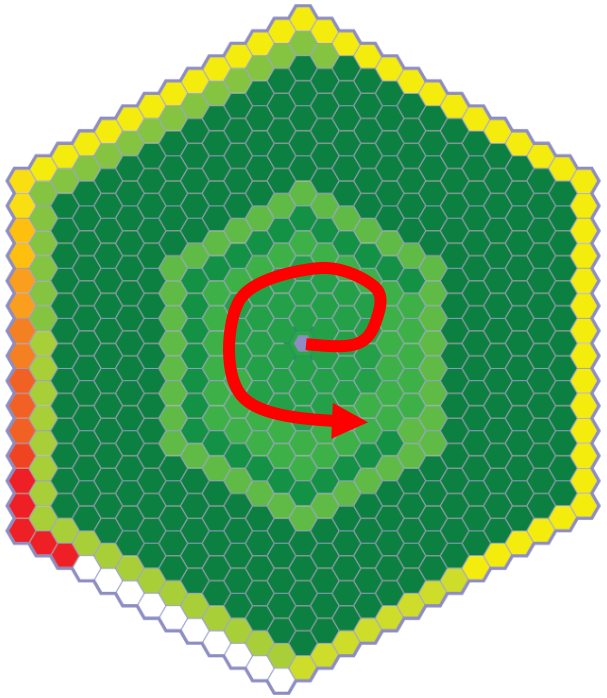
access importance: indomain multiple datasets
μ(σ) = 0.89 (0.17)[n=82, 2]

EarthCube Active



access importance: indomain multiple datasets
μ(σ) = 0.9 (0.17)[n=104, 6]

All Others

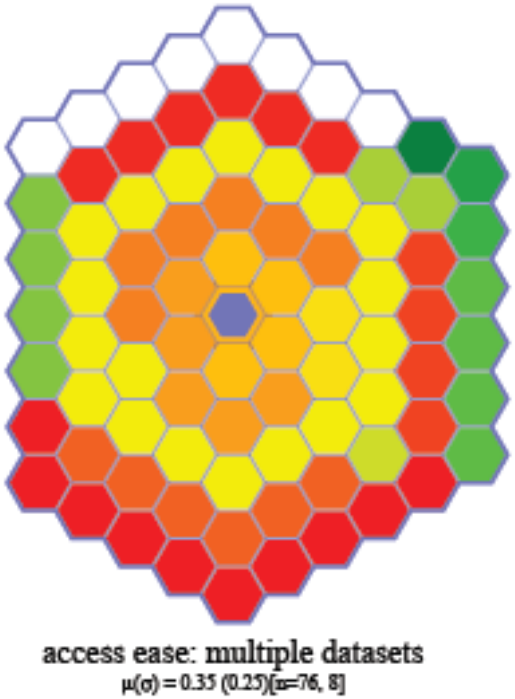


access importance: indomain multiple datasets
μ(σ) = 0.86 (0.2)[n=549, 10]

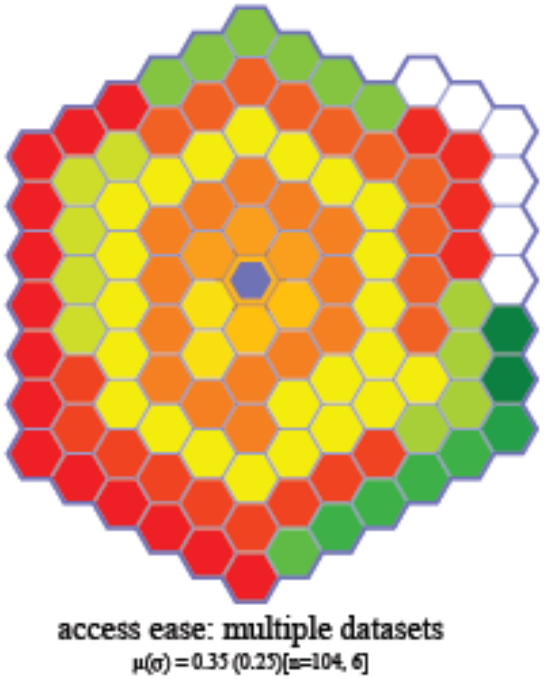
Comment: Vast majority are extremely positive on importance, with just a handful who are neutral or negative.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models in your field or discipline?	.89 (.17)	.90 (.17)	.86 (.20)

EASE of integrating multiple datasets, observations, visualization tools, and/or models in your field or discipline?

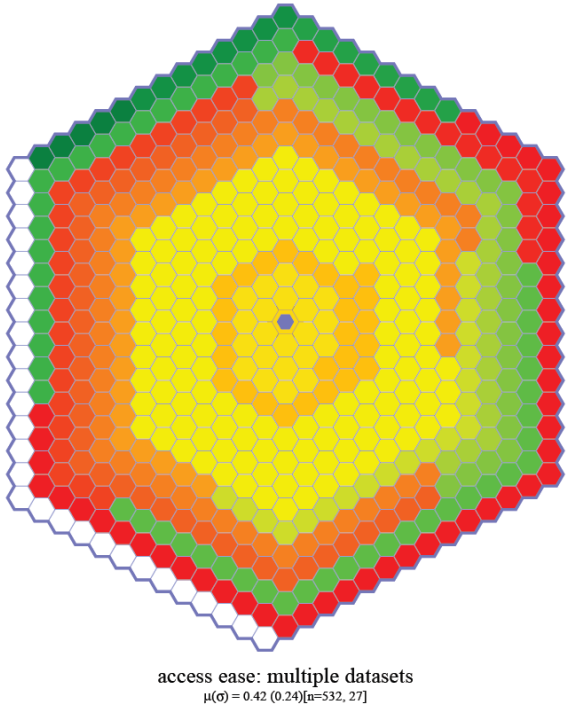
Early Career



EarthCube Active



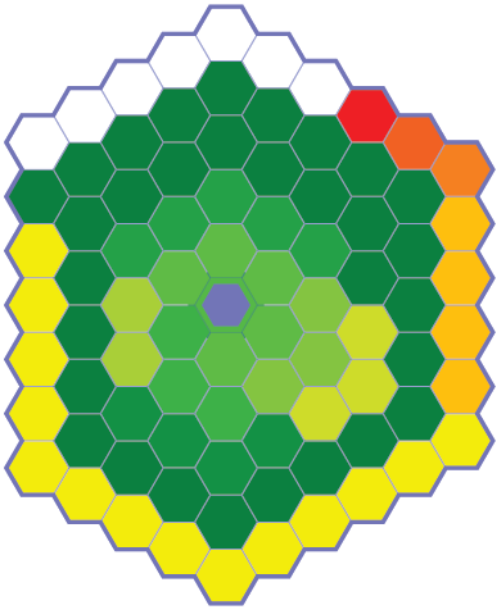
All Others



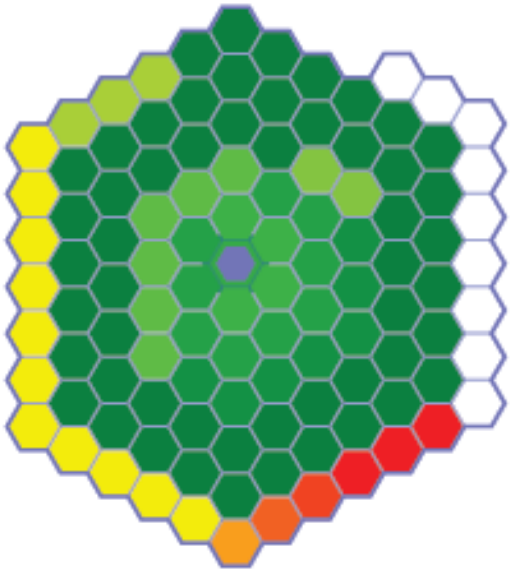
Comment: Vast majority of respondents are neutral or negative, with many strongly negative; early career and EC active are most negative.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
How EASY is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models in your field or discipline?	.35 (.25)	.35 (.25)	.42 (.24)

IMPORTANCE integrating multiple datasets, observations, visualization tools, and/or models spanning fields/disciplines?

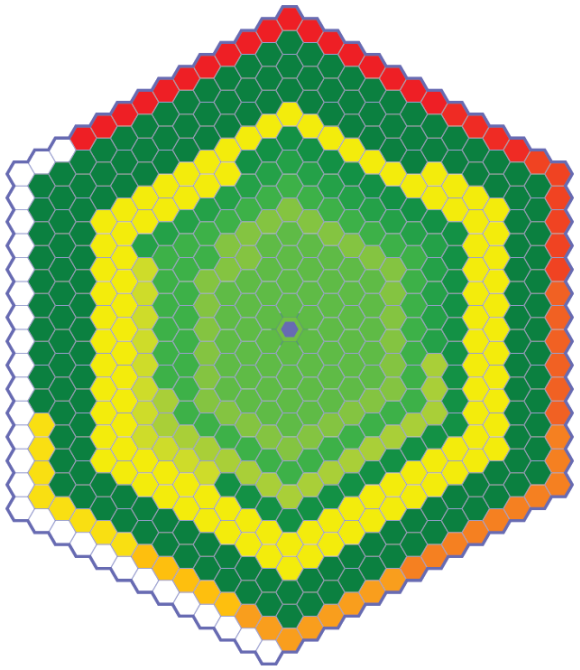
Early Career



EarthCube Active



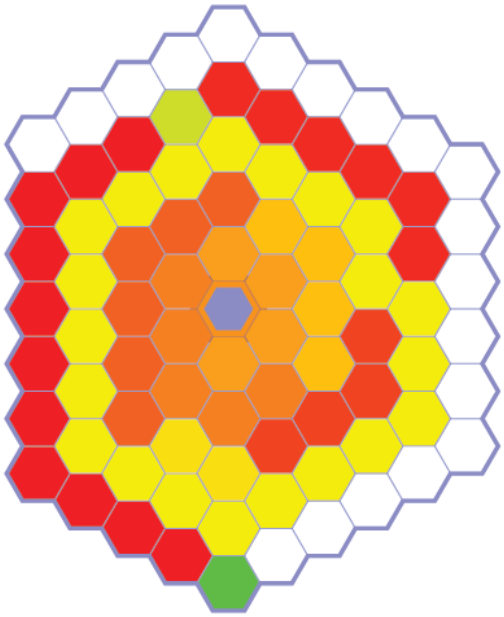
All Others



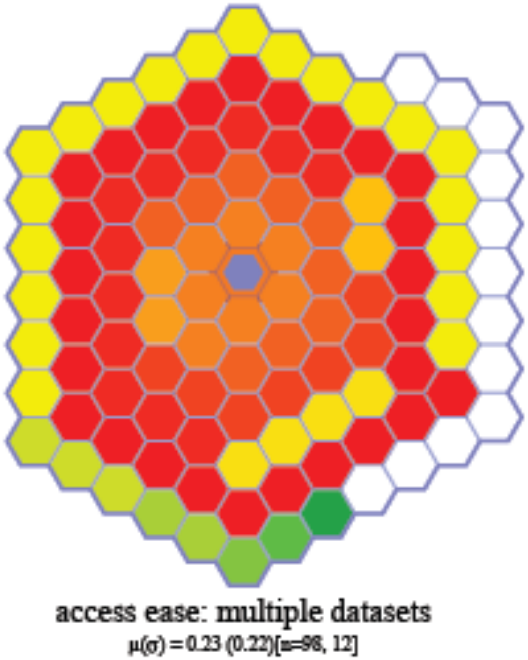
Comment: EC active are most positive, though most in all groups are positive (with some neutral and negative outliers).	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
How IMPORTANT is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models that span different fields or disciplines?	**.76 (.25)**	**.83 (.24)**	**.71 (.27)**

EASE of integrating multiple datasets, observations, visualization tools, and/or models that span different fields or disciplines?

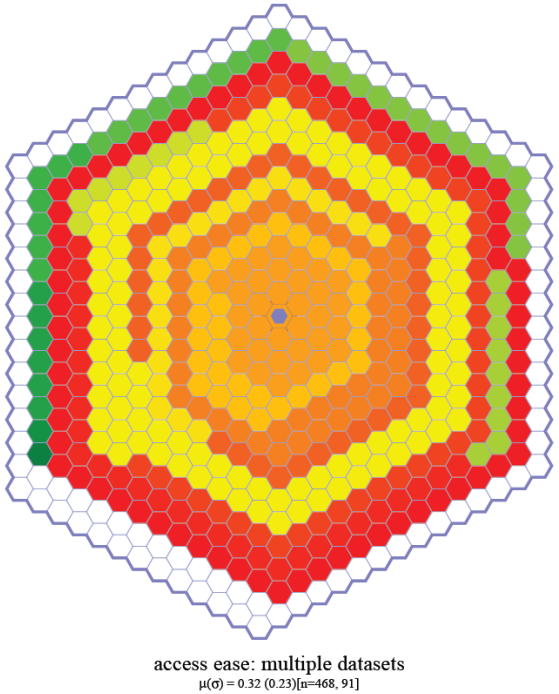
Early Career



EarthCube Active



All Others



Comment: EC active are most negative, though all respondents report difficulty in access (with only a handful of positive outliers).	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
How EASY is it for you to find, access, and/or integrate multiple datasets, observations, visualization tools, and/or models that span different fields or disciplines?	**.28 (.20)**	**.23 (.22)**	**.32 (.23)**

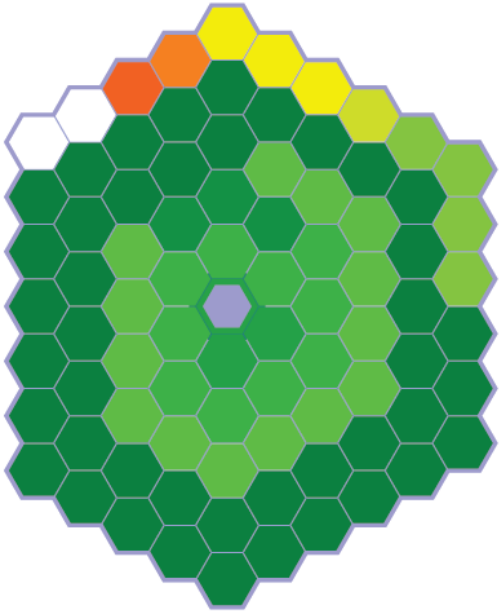
Top twenty-five cited sources of data – Consider the metadata implications. . .

1. **NOAA (NODC, NGDC, NDBC, NCEP, etc.))** (17%)
2. **NASA (JPL, ESA, etc.)** (11%)
3. **Colleagues/Clients** (10%)
4. **The web (unspecified)** (9%)
5. **NCAR, UCAR, Unidata** (8%)
6. **Publications** (8%)
7. **USGS** (8%)
8. **IEDA (GeoRock, EarthChem, MGDS, etc.)** (8%)
9. **State and local government** (3%)
10. **International (PANGEA, etc.)** (2%)
11. **DOE** (2%)
12. **EPA** (1%)
13. **Google/Google Earth** (1%)
14. **NSIDC, NIC** (1%)
15. **USDA** (1%)
16. **IRIS, EarthScope** (1%)
17. **Neotoma, PBDB, Macrostrat** (1%)
18. **BCO-DMO, JGOFS, WOCE, CLIVAAR, Geotraces** (1%)
19. **IODP** (1%)
20. **DOD (Navy, Army, Army Corps of Engineers)** (1%)
21. **Open topography, NCALM** (1%)
22. **LTER** (1%)
23. **UNAVCO** (1%)
24. **MagIC** (under 1%)
25. **Private sector companies (IRI, ESRI)** (under 1%)

Note: All percentages rounded to the nearest whole number

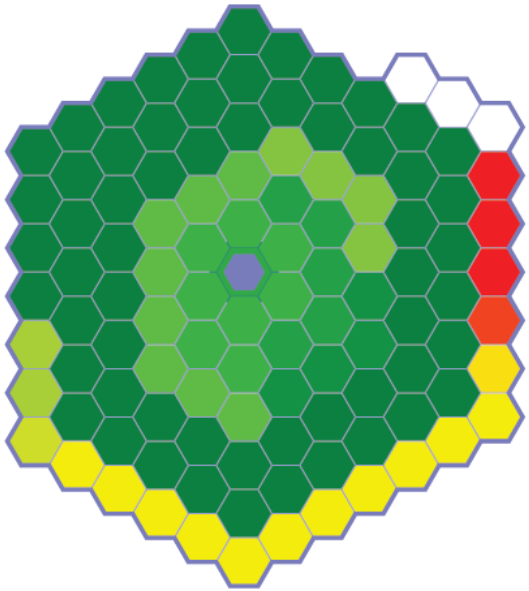
Specifying guidelines regarding geoscience data

Early Career



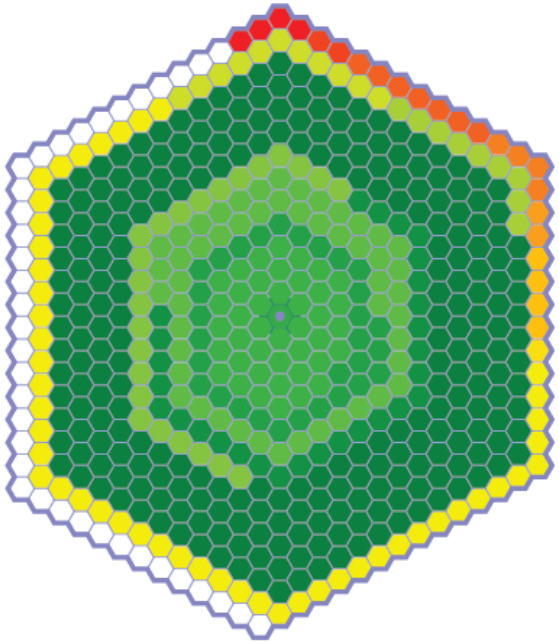
need guidelines: data sharing/access
μ(σ) = 0.87 (0.17)[n=82, 2]

EarthCube Active



need guidelines: data sharing/access
μ(σ) = 0.84 (0.23)[n=107, 3]

All Others

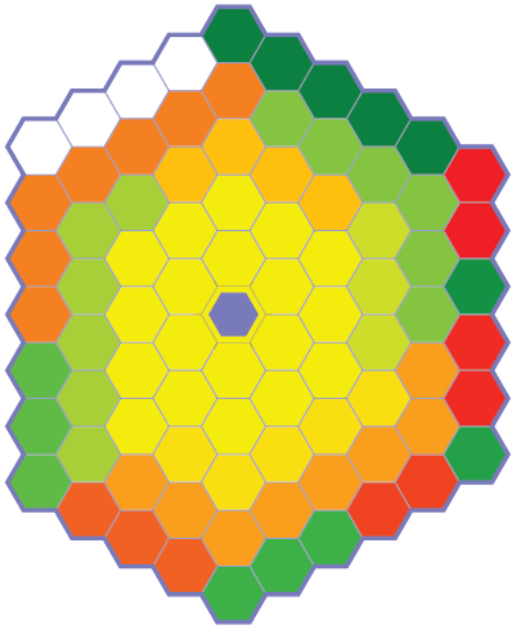


need guidelines: data sharing/access
μ(σ) = 0.83 (0.21)[n=522, 37]

Comment: A "pull" for guidelines/standards (strongest by early career), with a small number of neutral or negative responses.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
The EarthCube initiative should specify guidelines so there is more interoperability and uniformity in discovering, accessing, sharing, and disseminating geoscience data	.87 (.17)	.84 (.23)	.83 (.21)
The EarthCube initiative should specify guidelines so there is more interoperability and uniformity in geoscience visualization tools	.84 (.18)	.71 (.26)	.77 (.25)

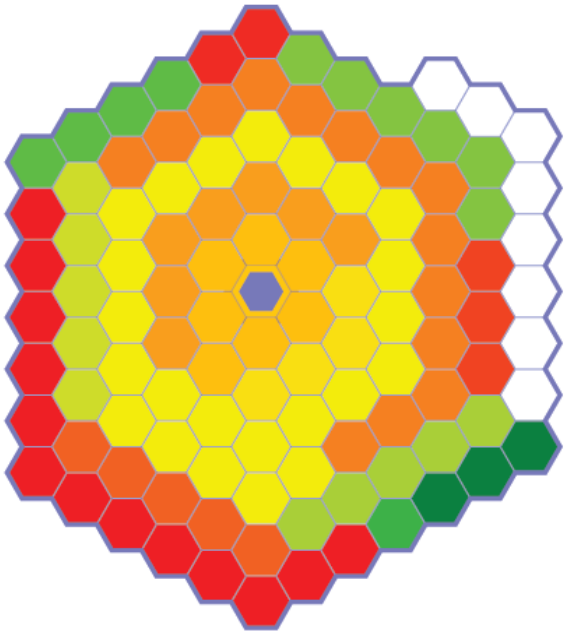
Adequacy of current suite of tools and modeling software

Early Career



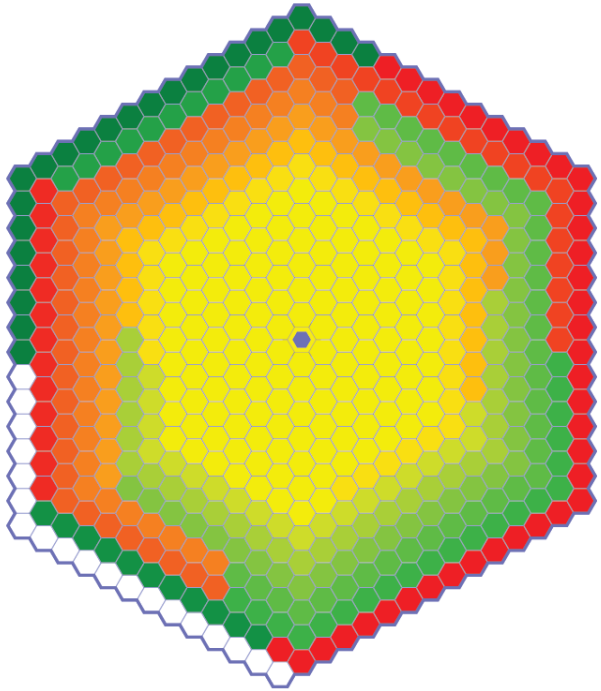
adequacy current data suite
μ(σ) = 0.49 (0.24)[n=80, 4]

EarthCube Active



adequacy current data suite
μ(σ) = 0.39 (0.24)[n=102, 8]

All Others



adequacy current data suite
μ(σ) = 0.47 (0.25)[n=540, 19]

Comment: Current suite seen as inadequate – motivation for EarthCube; most negative are EC active.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
Please use the scale ranging from Inadequate to Adequate to assess the present suite of publicly accessible datasets, data analysis tools, and modeling software – to what degree is it adequate for your research and education needs?	**.49 (.24)**	**.39 (.24)**	**.47 (.25)**

2. How can we incentivize researchers and providers to curate their data, organize it with useful metadata, and make it publicly available?

Lewin's force field analysis

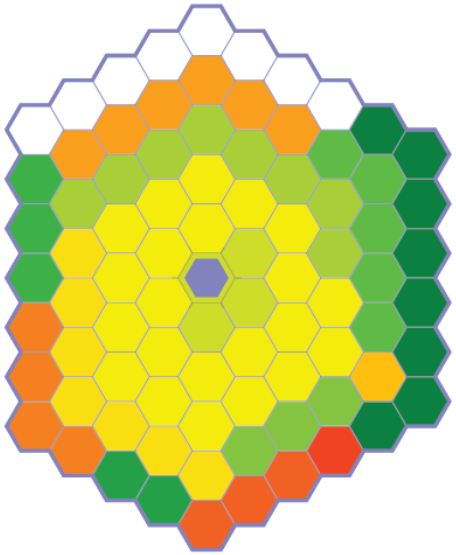
Interdisciplinary Innovation in the Geosciences

Urgency of geoscience research	Technical barriers to interoperability
Engaging research questions	Career development complications
Pos. signals from funding agencies	Publication/translation challenges
Strategic priorities of universities	"Birds of a feather" tendencies
Colleagues open to collaboration	Risk aversion tendencies

3. Maximum impact of data occurs when analytics make use of all available relevant data; how can analytics developers be challenged to make this standard practice?

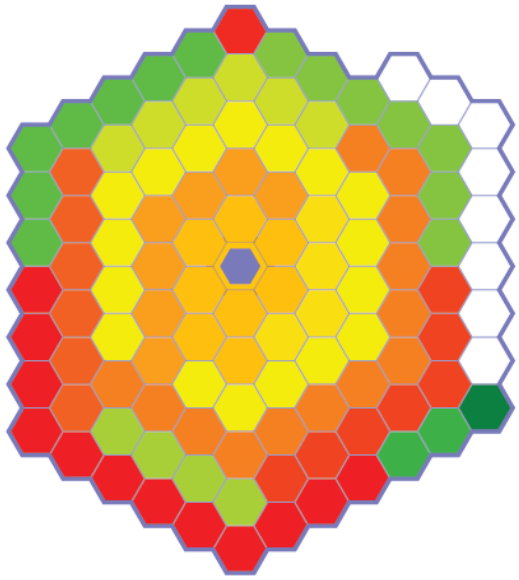
Cooperation and sharing of data, software among Geo

Early Career



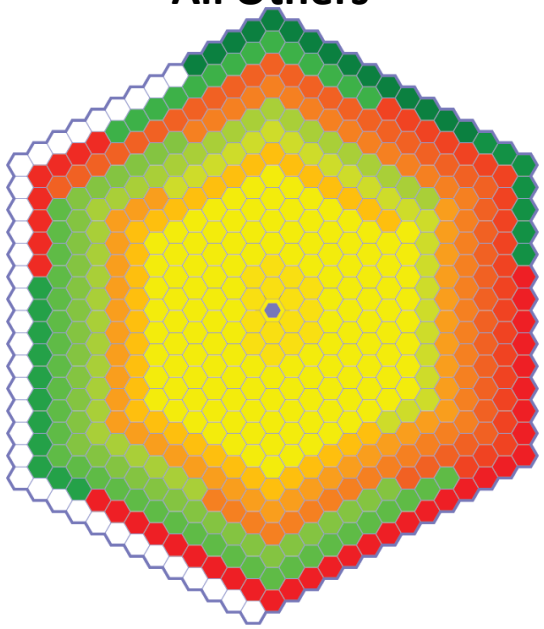
current: coop./sharing among geoscientists
μ(σ) = 0.56 (0.22)[n=76, 8]

EarthCube Active



current: coop./sharing among geoscientists
μ(σ) = 0.38 (0.23)[n=102, 8]

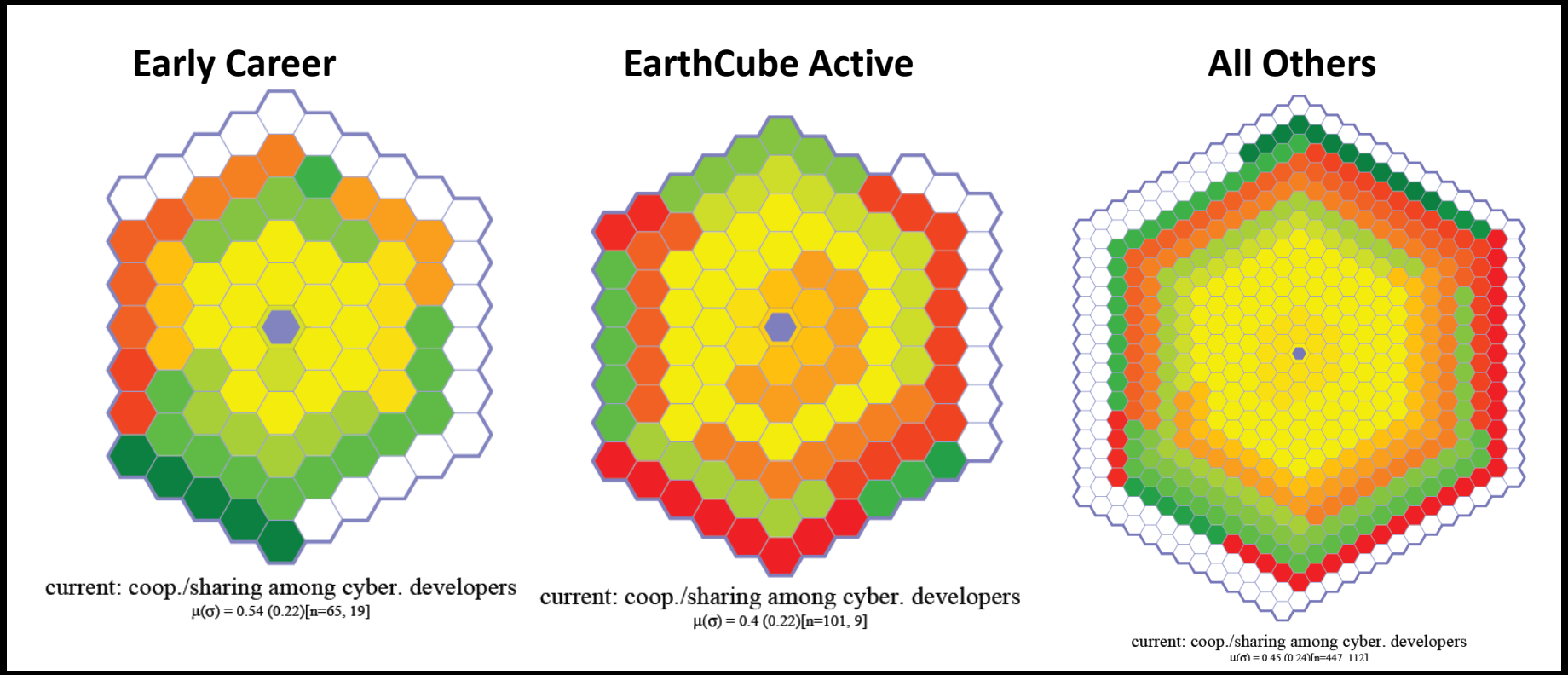
All Others



current: coop./sharing among geoscientists
μ(σ) = 0.45 (0.24)[n=524, 35]

Comment: Strong negative views on cooperation and sharing in Geo community; handful of "bright spots;" most concerned views among EC active.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
There is currently a high degree of cooperation and sharing of data, models, and simulations among geoscientists	**.56 (.22)**	**.38 (.23)**	**.45 (.24)**

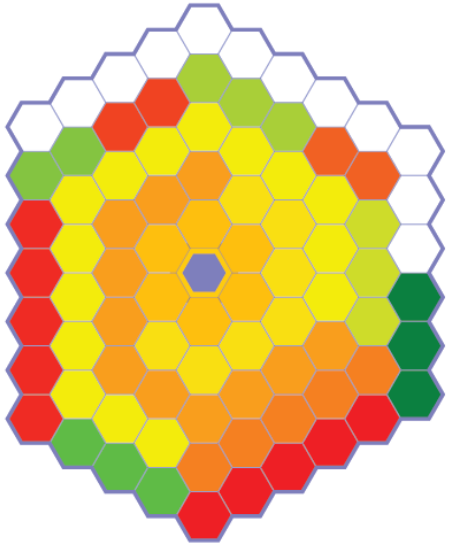
Cooperation and sharing of data, software within Cyber community



Comment: Strong negative views on cooperation and sharing in Cyber community; handful of "bright spots;" EC active most concerned.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
There is currently a high degree of cooperation and sharing of software, middleware and hardware among those developing and supporting cyberinfrastructure for the geosciences	**.54 (.22)**	**.40 (.22)**	**.45 (.24)**

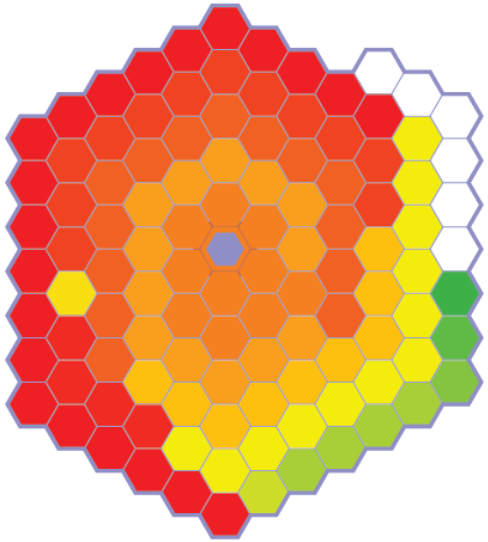
Communication and Collaboration: Geo and Cyber

Early Career



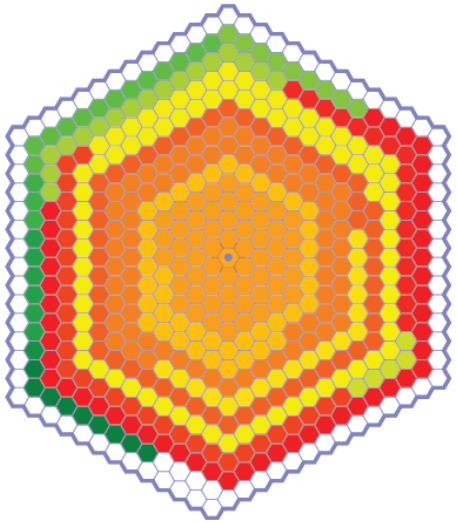
current: sufficient collab. between Geo and Cyber
μ(σ) = 0.39 (0.23)[n=72, 12]

EarthCube Active



current: sufficient collab. between Geo and Cyber
μ(σ) = 0.25 (0.19)[n=104, 6]

All Others

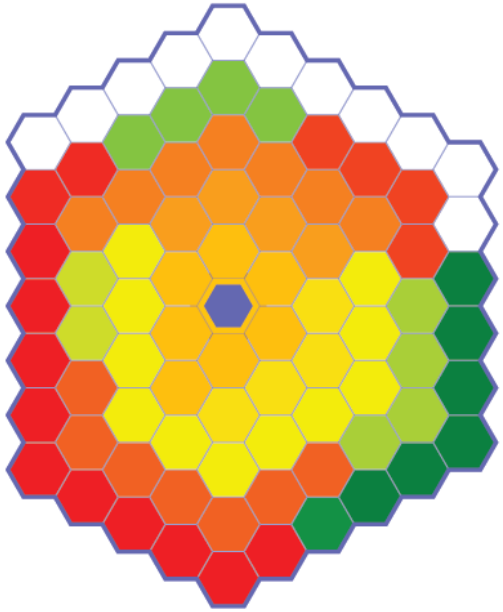


current: sufficient collab. between Geo and Cyber
μ(σ) = 0.33 (0.22)[n=476, 83]

Comment: Major concerns with communication and collaboration between Geo and Cyber communities; strongest concerns among EC active.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
There is currently sufficient communication and collaboration between geoscientists and those who develop cyberinfrastructure tools and approaches to advance the geosciences	.39 (.23)	.25 (.19)	.33 (.22)

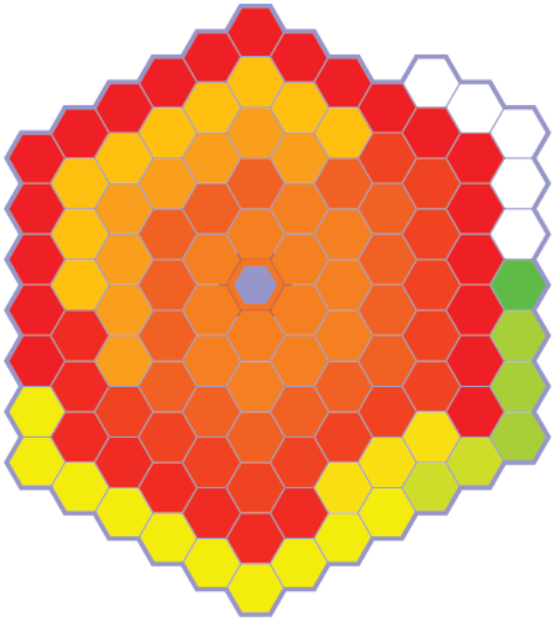
Geo and Cyber – End-user training

Early Career



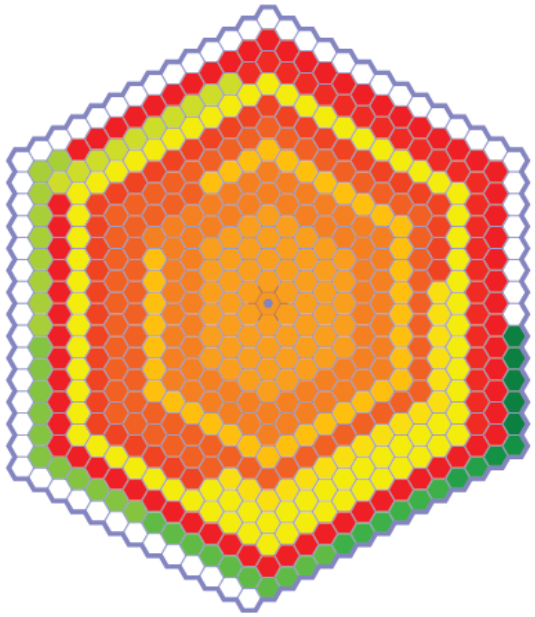
current: sufficient end-user training
μ(σ) = 0.38 (0.28)[n=73, 11]

EarthCube Active



current: sufficient end-user training
μ(σ) = 0.22 (0.18)[n=105, 5]

All Others

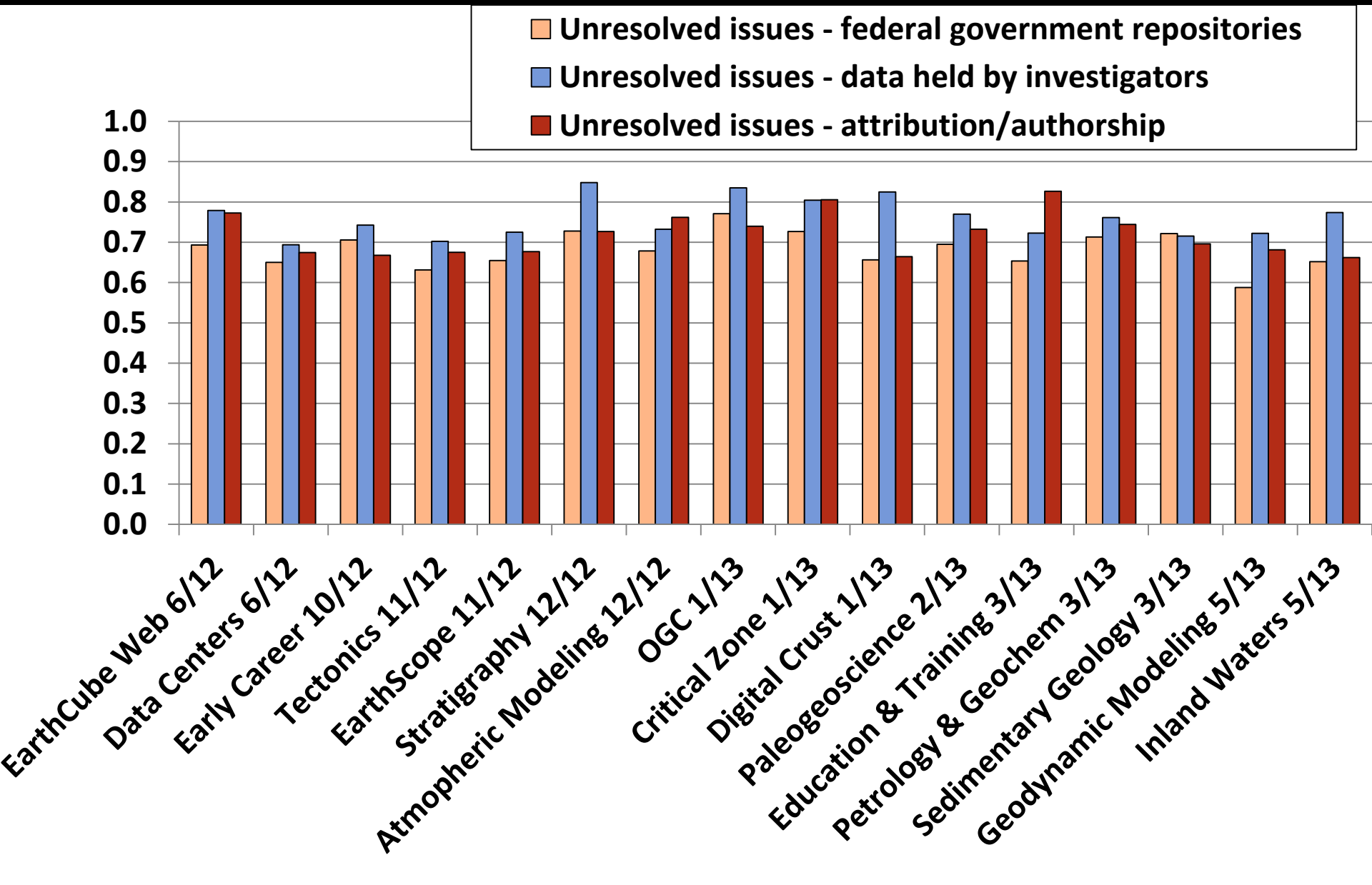


current: sufficient end-user training
μ(σ) = 0.3 (0.21)[n=499, 60]

Comment: Major concerns end-user knowledge of Cyber by Geo, with strongest concerns among EC active.	Early Career μ (α)	EC Active μ (α)	All Others μ (α)
There is currently sufficient geoscience end-user knowledge and training so they can effectively use the present suite of cyberinfrastructure tools and train their students/colleagues in its use	**.38 (.28)**	**.22 (.18)**	**.30 (.21)**

4. What are the data ownership and personal identifiable information issues (obstacles/solutions) that can be addressed in this context?

Unresolved issues. . .



Top Ten Barriers to Sharing Data (categories):

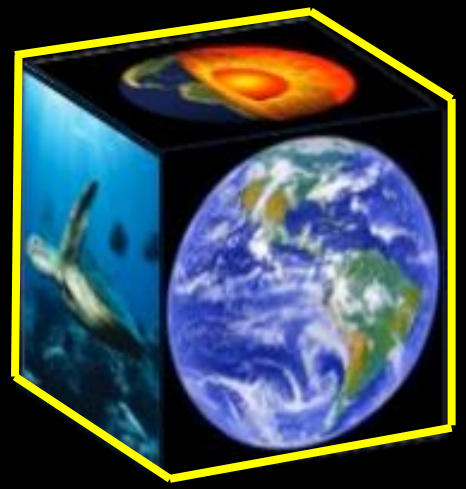
1. **No time/Needs too much QA/QC**
2. **No repository/No known repository**
3. **Inadequate standards/No standardized formats**
4. **Want to publish first/Don't want to be scooped**
5. **File size too large/Server size too small**
6. **Classified/proprietary/Agency or company restrictions**
7. **No credit/No incentive to share**
8. **Cost**
9. **Not sure what to do**
10. **Not sure anyone wants it**

Note: Approximately 45% of respondents did not respond to the open ended question "It is difficult to share my data because. . . " and another 6% said it was easy to share their data. The balance of responses were organized into the above categories; some individuals cited more than one reason (all of which were tabulated).

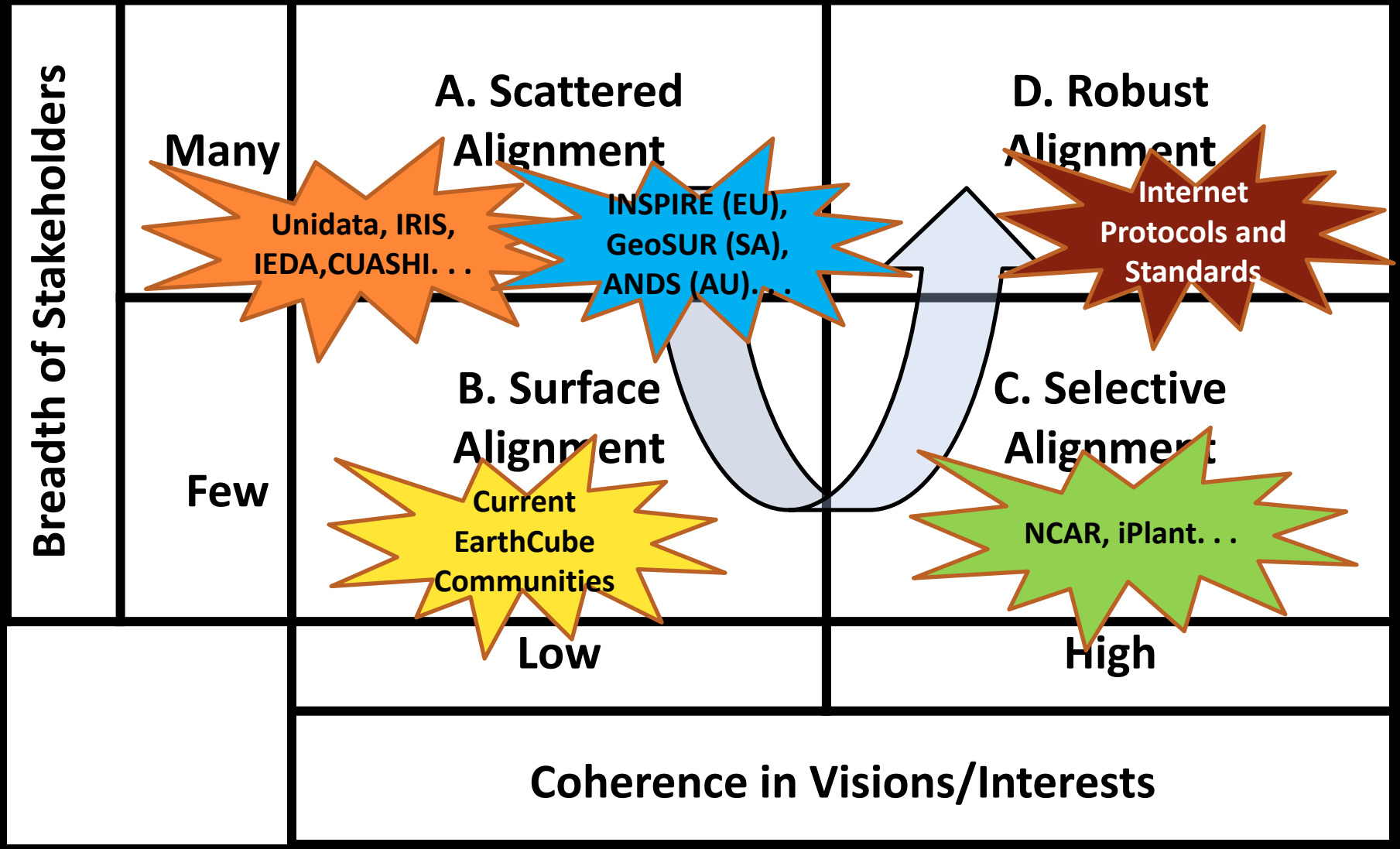
5. What are the top two data/metadata problems you would like to solve?

Problems 1 & 2: Forging a Robust and Agile Institutional Context for Data

- Building Blocks (6/13-present)
- Governance (6/13-present)
- Domain Workshops (10/12-present)
- Special Interest Groups (6/12)
- Charrette II (6/12)
- Stakeholder Alignment Survey (3/12)
- Roadmaps (3/12-8/12)
- Initial Groups (3/12)
- Expressions of Interest (1/12-4/12)
- Charrette I (11/11)
- White Papers (10/11)



Forms of alignment



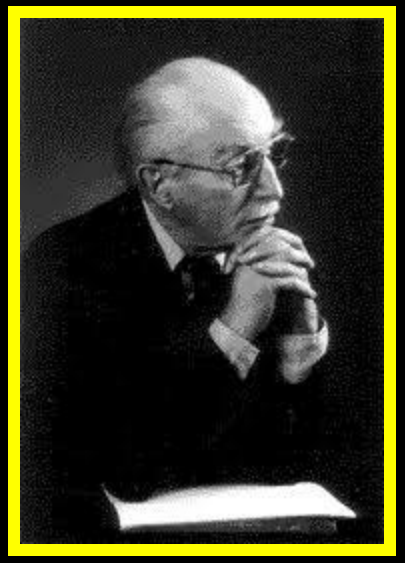
Today's most troubling and daunting problems have common features: some of them arise from human numbers and resource exploitation; they require long-term commitments from separate sectors of society and diverse disciplines to solve; simple, unidimensional solutions are unlikely; and failure to solve them can lead to disasters.

In some ways, the scales and complexities of our current and future problems are unprecedented, and it is likely that solutions will have to be iterative . . .

Institutions can enable the ideas and energies of individuals to have more impact and to sustain efforts in ways that individuals cannot.

From "Science to Sustain Society," by Ralph J. Cicerone, President, National Academy of Sciences, 149th Annual Meeting of the Academy (2012)

Appendix



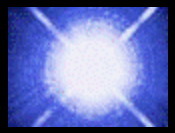
Looking ahead . . .

". . . We are moving towards another type of society than that to which we have become accustomed. This is sometimes referred to as a new service society, the society of the second industrial revolution or the post-industrial society. There is no guarantee of our safe arrival. Not only are the interdependencies greater – they are differently structured. . . [and] demand a new mobilization of the sciences."

– Source: Eric L. Trist, from paper on "Social Aspects of Science Policy" (March, 1969) cited in *Towards a Social Ecology: Contextual Appreciation of the Future in the Present* by Fred E. Emery and Eric L. Trist (London: Plenum Press, 1973)

Most important challenges of the 21st Century, as identified by NAE


Make solar energy economical


Provide energy from fusion


Develop carbon sequestration methods


Manage the nitrogen cycle


Provide access to clean water


Restore and improve urban infrastructure


Advance health informatics


Engineer better medicines


Reverse-engineer the brain


Prevent nuclear terror


Secure cyberspace


Enhance virtual reality


Advance personalized learning


Engineer the tools of scientific discovery

Source: http://www.engineeringchallenges.org/

Institutional and systems requirements

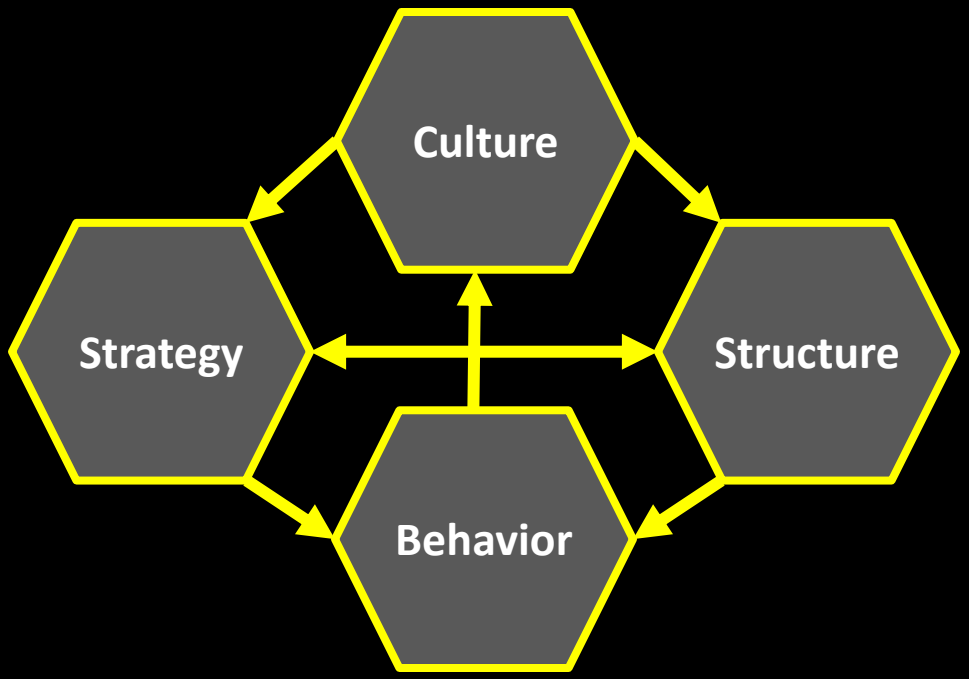
Creating Value — *. . . expanding the "pie" and enabling systems transformation*

Mitigating Harm — *. . . anticipating and mitigating externalities and catastrophic systems failures*

Defining stakeholder alignment. . .

"The extent to which interdependent stakeholders orient and interact with one another to advance their separate and shared interests."

A simplified conceptual framework . . .



Steps in the process

Phase I: Navigator (1.0)

1.1 Define scale and scope
1.2 Form launch team
1.3 Plan launch events

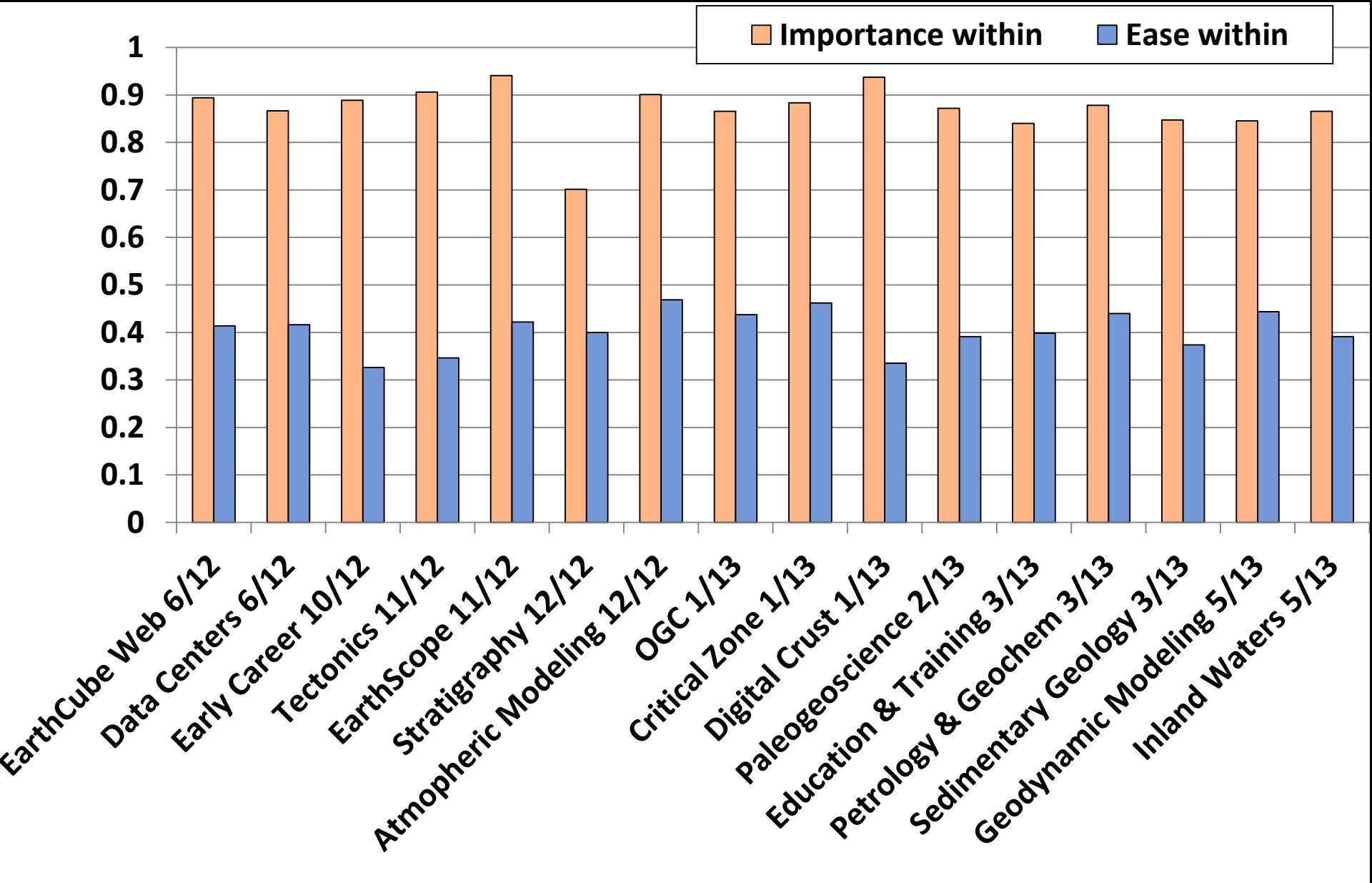
Phase II: Map (2.0)

2.1 Specify stakeholders
2.2 Identify interests
2.3 Develop instrument(s)
2.4 Survey representative sample
2.5 Visualize alignment/ misalignment

Phase III: Journey (3.0)

3.1 Construct shared vision of success (future state)
3.2 Assess strengths and weakness (current state)
3.3 Align resources and support systems (delta state)
3.4 Charter appropriate forums (delta state)
3.5 Establish milestones and metrics (delta state)
3.6 Address misaligned incentives (delta state)
3.7 Ensure internal alignment (delta state)
3.8 Manage leadership transitions (delta state)
3.9 Check and adjust (new current state and new future state)

Integrating multiple datasets, observations, visualization tools, and/or models – within



Integrating multiple datasets, observations, visualization tools, and/or models – across